Docket No.: POU920040002US1

Inventor: Ferri et al.

Title: FACILITATING ALLOCATION OF RESOURCES IN A HETEROGENEOUS COMPUTING ENVIRONMENT

APPLICATION FOR UNITED STATES

LETTERS PATENT

"Express Mail" Mailing Label No.: EL965409108US
Date of Deposit: 3-10-2004

I hereby certify that this paper is being deposited with the United States Postal Service as "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to: Mail Stop PATENT APPLICATION, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.

Name: Susan L. Nelson

Signature: Susan L Nelson

INTERNATIONAL BUSINESS MACHINES CORPORATION

FACILITATING ALLOCATION OF RESOURCES IN A HETEROGENEOUS COMPUTING ENVIRONMENT

Cross-Reference to Related Applications

[0001] This application contains subject matter which is related to the subject matter of the following application, which is assigned to the same assignee as this application and hereby incorporated herein by reference in its entirety:

[0002] "MANAGING PROCESSING WITHIN COMPUTING ENVIRONMENTS INCLUDING INITIATION OF VIRTUAL MACHINES," Bulson et al., (IBM Docket No. POU920030140US1), Serial No. 10/667,163, filed September 17, 2003.

Technical Field

[0003] This invention relates, in general, to heterogeneous computing environments, and in particular, to facilitating the allocation of resources in a heterogeneous computing environment.

Background of the Invention

[0004] A heterogeneous computing environment includes a plurality of nodes having different architectures and operating systems. For instance, at least one node of the environment is based on a different architecture executing a different operating system than at least one other node of the environment. One example of a heterogeneous computing environment is a grid computing environment.

[0005] A grid computing environment includes a number of nodes, such as workstations or servers, that cooperate to solve large computing problems. Typically, a grid is spread out over a large geographic area. Individual nodes that contribute to the grid may have many other alternate purposes – at various times, they may be used as nodes within a cluster or as individual workstations, as well as grid members. Typically,

the decision to make an individual node or group of nodes available to the grid is based on a number of items, including how busy the node is in its non-grid roles, the demand for grid nodes, and the types of resources dedicated to the node. These resources, such as storage, memory, compute power and file system resources, can be allocated to the greater grid in order to create a powerful, dynamic, problem solving environment.

[0006] In the current state of the art, there is a strong affinity between the grid node and the characteristics of the job to be executed on that node. In particular, the node selected to run the job or a portion thereof is to have the correct version of the operating system and the correct platform architecture (i.e., the correct environment). This is a result of the job having bound executable files that are compiled and linked for a particular operating system on a particular platform. For example, the executable files contain machine level instructions which can only run on machines of the same environment. This shortcoming restricts the grid to allocate resources of only those nodes having the same environment as the job to be executed.

[0007] Previously, attempts have been made to overcome this shortcoming. However, a need still exists for a capability that facilitates allocation of resources in a heterogeneous computing environment.

Summary of the Invention

[0008] The shortcomings of the prior art are overcome and additional advantages are provided through the provision of a method of facilitating allocation of resources in a heterogeneous computing environment. The method includes, for instance, obtaining, by a resource manager of the heterogeneous computing environment, one or more attributes relating to one or more nodes coupled to the resource manager, the one or more attributes specifying at least one compatible environment supported by the one or more nodes; and taking into consideration, by the resource manager, at least one attribute of the one or more attributes in allocating one or more resources of at least one node of the one or more nodes to a request.

[0009] System and computer program products corresponding to the above-summarized method are also described and claimed herein.

[0010] Additional features and advantages are realized through the techniques of the present invention. Other embodiments and aspects of the invention are described in detail herein and are considered a part of the claimed invention.

Brief Description of the Drawings

[0011] The subject matter which is regarded as the invention is particularly pointed out and distinctly claimed in the claims at the conclusion of the specification. The foregoing and other objects, features, and advantages of the invention are apparent from the following detailed description taken in conjunction with the accompanying drawings in which:

[0012] FIG. 1 depicts one embodiment of a heterogeneous computing environment incorporating and using one or more aspects of the present invention;

[0013] FIG. 2 depicts one example of a cluster that may be employed in the heterogeneous computing environment of FIG. 1, in accordance with an aspect of the present invention;

[0014] FIG. 3 depicts one example of a plurality of clusters of a grid computing environment, in accordance with an aspect of the present invention;

[0015] FIG. 4 depicts one embodiment of a grid computing environment in which the resources available to the grid are restricted;

[0016] FIG. 5 depicts one embodiment of the logic associated with providing additional resources to the grid, in accordance with an aspect of the present invention; and

[0017] FIG. 6 depicts one embodiment of a grid computing environment in which additional resources are available to the grid for allocation, in accordance with an aspect of the present invention.

Best Mode for Carrying Out the Invention

[0018] In accordance with an aspect of the present invention, a capability is provided to facilitate allocation of resources in a heterogeneous computing environment. The heterogeneous environment includes at least one resource manager that is responsible for determining which nodes of the environment can process a particular request. To make this determination, previously, the resource manager would consider those nodes having the same environment (i.e., architecture and operating system) as the request. Further, as an enhancement, the resource manager could also consider nodes of different generations of the same architecture as the request, as described in co-pending patent application Serial No. 10/667,163, entitled "Managing Processing Within Computing Environments Including Initiation Of Virtual Machines," filed September 17, 2003.

[0019] With one or more aspects of the present invention, the scope of nodes to be considered to process a particular request is greatly expanded. For example, nodes can be considered that have different native environments than the request. These heterogeneous nodes can be considered, since they are capable of supporting other environments, although their native environments are different than the request.

[0020] The resource manager of the heterogeneous computing environment obtains information from the various nodes in the environment and uses that information to determine which nodes can be used to process a request. The information obtained by the resource manager includes attributes relating to the one or more environments (e.g., platforms and operating systems) that are supported by but not native to the nodes. These compatibility attributes are made available to the resource manager to broaden the scope of nodes, and therefore, resources, available to process a specific request.

[0021] One example of a heterogeneous computing environment is depicted in FIG. 1. In this example, the heterogeneous environment is a grid computing environment 100 including, for instance, a plurality of user workstations 102 (e.g., laptops, notebooks, such as ThinkPads, personal computers, RS/6000's, etc.) coupled to a job management service 104 via, for instance, the internet, extranet, or intranet. Job management service 104 includes, for instance, a web application to be executed on a web application server, such as Websphere offered by IBM®, or distributed across a plurality of servers. It has the responsibility for accepting user requests and passing the requests to the appropriate nodes of the environment. As one example, a user interacts with the job management service via a client application, such as a web browser or a standalone application. There are various products that include a job management service, including, for instance, LSF offered by Platform (www.platform.com), and Maui, an open source scheduler available at http://www.supercluster.org. (IBM® is a registered trademark of International Business Machines Corporation, Armonk, New York, U.S.A. Other names used herein may be registered trademarks, trademarks or product names of International Business Machines Corporation or other companies.)

[0022] Job management service 104 is further coupled via the internet, extranet or intranet to one or more data centers 106. Each data center includes, for instance, one or more nodes 108, such as mainframes, workstations and/or servers. The nodes of the environment are heterogeneous nodes in that at least one node is based on a different architecture and/or is executing a different operating system than at least one other node. For example, one node may be based on the x86 architecture running a Linux operating system and another node may be based on the PowerPC architecture running AIX.

[0023] As described herein, a grid computing environment includes a plurality of nodes which cooperate to solve a large computing problem. Individual nodes that contribute to the grid may have many other alternate purposes. For example, nodes within the grid may be used as individual workstations or form a cluster, an example of which is depicted in FIG. 2. A cluster 200 includes, for instance, a set of homogeneous

nodes 202, which are managed by a resource manager 204, such as a cluster resource manager. The cluster resource manager receives requests from users and is responsible for allocating cluster resources to the individual requests. The manner in which the resources are allocated depends on the requirements (e.g., storage, CPU requirements, etc.) of the request. It is the responsibility of the cluster resource manager to use the resources efficiently to maximize throughput. An example of a cluster resource manager is Loadleveler, offered by International Business Machines Corporation, Armonk, New York.

[0024] One or more clusters may be coupled together to form a grid computing environment. In a grid environment, an additional layer of software is added which allocates nodes to the greater grid, as shown in FIG. 3. A grid computing environment 300 includes a job management service, such as a grid resource manager 302, which is coupled to a plurality of cluster resource managers 304 of a plurality of clusters. A request is submitted to the grid resource manager, and the grid resource manager is aware of one or more sets of nodes, each of which may have different architectures and operating systems. It is from this pool of architectures and operating systems that the grid resource manager draws on when evaluating available resources for a grid request. For example, if a grid request (e.g., a job or a portion thereof) is submitted which is a Linux executable compiled for an x86 cluster, the grid manager only considers the set of nodes that exactly satisfy these conditions to select candidate resources. This shortcoming is illustrated in FIG. 4, wherein a large number of resources might stand idle because of the operating system and architecture constraints of the submitted request.

[0025] Referring to FIG. 4, a grid resource manager 400 only considers compute nodes 402 coupled to cluster resource manager 404, since the grid resource manager is aware that those nodes have the same environment as the executable to be run. That is, the Linux executable compiled for an x86 cluster can run on those nodes, since they are based on the x86 architecture and are executing the Linux operating system. Nodes 406, managed by a cluster resource manager 408, running an AIX operating system on a

PowerPC architecture are not considered for resource allocation, since they are not the same environment as the submitted executable.

[0026] However, in accordance with an aspect of the present invention, nodes 406 may be considered. That is, a capability is provided to enable a grid resource manager to consider additional nodes, if those nodes can support further operating systems and architectures. For example, it is possible that although one or more nodes of a grid are native to one environment (i.e., have a particular hardware platform and operating system), that they may also be able to support other environments. For instance, a node that has a native operating environment of PowerPC executing AIX may be able to support an environment of x86 running Linux. Thus, that node can process requests (e.g., execute programs) that require either environment, as indicated by the request either explicitly or implicitly.

[0027] In order for a request to be executed on a node that is not native to the request, the request is moved to the other environment. For example, if the request is to run a program, the program, which is written for the native environment, is moved to the other environment. Techniques exist for moving a program from one environment and making it executable in another environment. One such technique includes using an application programming interface (API). Programs access operating system routines, such as reading and writing files, through a series of well defined operating system functions defined in an API. A well constructed program written in a high level programming language, such as the C programming language, can easily be ported from one operating system to another if both operating systems support the same API. For example, a program written in C with a number of API calls might run on an AIX operating system, but could be ported to run on Linux, if Linux supports the same set of APIs that AIX does. The same could hold true in porting an application from an x86 (Intel 32 bit) platform to a 64 bit platform - if an API is provided, the program should port rather easily.

[0028] The drawback of APIs is that they are programming interfaces, and require a recompile of code to port the application from one environment to another. Again, going back to the example of a C program running in AIX, the same program would require a recompile in a Linux environment, using the Linux APIs, to make the program execute successfully under Linux. This concept of forcing a recompile for each program that is ported from one environment to another is time consuming, and sometimes, uncovers hidden flaws in the API. Thus, other techniques have been sought.

[0029] One technique, referred to as ABI or Application Binary Interface, provides a more straightforward technique for running a program in a different environment. ABI provides a technique of taking an executable from one environment and running it in another environment without recompilation through the use of an emulation software layer or through direct hardware support on a target machine. One architecture that uses ABI is the AMD64 architecture, offered by AMD, Sunnyvale, California. An example of ABI is described in "Binary Compatibility," http://gcc.gnu.org/onlinedocs/gcc/Compatibility.html, the content of which is hereby incorporated herein by reference in its entirety.

[0030] In accordance with an aspect of the present invention, nodes that are capable of supporting different architectures, such as those that are ABI capable, are exposed to the grid resource manager, so that the grid resource manager can use this information when allocating tasks to the individual clusters or nodes. This enables a wider group of nodes to become available to the greater grid.

[0031] On embodiment of the logic associated with exposing the different capabilities to the grid resource manager, which is then capable of using this information in its resource allocation, is described with reference to FIG. 5.

[0032] When a node, such as a workstation, comes online, STEP 500, it provides a set of attributes to its resource manager, such as the cluster resource manager, STEP 502. These attributes include, as examples, the platform (architecture) of the node; the

operating system and operating system level of the node; as well as a set of compatibility attributes, including, for instance, any additional operating systems and/or any additional platforms (architectures) supported by the node through, for instance, ABI. This information is provided to the cluster resource manager via, for instance, Web Services Calls or Web Services Notifications. For instance, an XML document specifying these attributes is transferred to the cluster resource manager using Web Services Calls or Notifications, such as a SOAP call. One example of SOAP is described in "SOAP Version 1.2 Part 0: Primer," Nilo Mitra, http://www.w3.org/TR/2003/REC-soap12-part0-20030624/, the content of which is hereby incorporated herein by reference in its entirety.

[0033] The cluster resource manager receives this information and percolates at least the compatibility attributes of the node (e.g., one or more non-native environments that it supports) to the grid resource manager, STEP 504. In one example, this information is provided to the grid resource manager using a similar mechanism as used to forward the information from a node to the cluster resource manager. For instance, the compatibility attributes are provided via a Web Services Call, such as a SOAP call, to a Web Service exposed by the grid resource manager. Although an example for providing the information is given herein, many other communications media are possible. The grid resource manager then takes these attributes into consideration when allocating resources to a request, STEP 506.

[0034] FIG. 6 depicts a pictorial illustration of the grid resource manager obtaining the compatibility attributes, such that the grid resource manager can use those attributes in allocating resources. As shown in FIG. 6, compute nodes 600 each have a native environment of AIX and PowerPC, but are x86 and Linux ABI compatible. The nodes provide this information to their cluster resource manager 602. The cluster resource manager then percolates this information up to grid resource manager 604. Thus, when the grid resource manager receives a request for a Linux executable, it can consider nodes 606, as well as nodes 600 when determining how to allocate the resources. For example, it can send queries to managers on nodes 600 and 606 to see if those nodes have the

needed resources. If one or more of the nodes have the resources, then at least one of the nodes is selected to process the request.

[0035] Described in detail above is a capability for facilitating allocation of resources in a heterogeneous computing environment. A resource manager, such as a grid resource manager, obtains information that identifies which nodes in the heterogeneous environment are able to support additional operating systems and platforms. The resource manager then uses this information to determine how to allocate resources. This advantageously expands the number of resources available to the heterogeneous computing environment for a particular request.

[0036] Although one example of a heterogeneous computing environment is described herein, many variations to the environment are possible without departing from the spirit of one or more aspects of the present invention. For instance, environments other than grid environments can incorporate and use one or more aspects of the present invention. As an example, any heterogeneous environment that uses an entity, such as a resource manager, to allocate resources can benefit from one or more aspects of the present invention. For instance, clusters having heterogeneous nodes, as well as other environments in which a manager manages a group of heterogeneous nodes, can use one or more aspects of the present invention. As a further example, the grid resource manager and the cluster resource manager are just examples of managers that can be used in accordance with one or more aspects of the present invention. Other types of managers can be used. Further, although examples of grid and cluster resource managers are provided, other alternatives exist. For instance, other applications or processes can be used.

[0037] As further examples, nodes can be different classes than that described herein (e.g., other than mainframes, workstations or servers) and/or can support different environments. Many environments (e.g., architectures and/or operating systems) are capable of using one or more aspects of the present invention. Further, a heterogeneous environment that includes a node having the same architecture but a different generation

as another node can incorporate one or more aspects of the present invention. Additionally, various types of interfaces, other than ABI, may also be used to move jobs to different environments. Moreover, various mechanisms may be used to percolate the compatibility attributes from the nodes to the grid resource manager or other manager.

[0038] As yet another example, the user can be replaced by an automated service or program. Further, a single request or job may include multiple jobs that run simultaneously on multiple nodes. This is accomplished similarly to that described above. For instance, the grid resource manager contacts a plurality of cluster managers and has those managers manage the plurality of requests. Many other variations also exist. As a further example, the environment may include one or more nodes that are partitioned. As yet a further example, one or more aspects of the present invention apply to Plug Compatible Machines (PCM) from Hitachi. Other examples are also possible.

[0039] Despite the type of environment, advantageously, one or more aspects of the present invention enable the harnessing of unutilized compute power which provides immediate economic benefits to an organization that has a large installed base of nodes.

[0040] The capabilities of one or more aspects of the present invention can be implemented in software, firmware, hardware or some combination thereof.

[0041] One or more aspects of the present invention can be included in an article of manufacture (e.g., one or more computer program products) having, for instance, computer usable media. The media has therein, for instance, computer readable program code means or logic (e.g., instructions, code, commands, etc.) to provide and facilitate the capabilities of the present invention. The article of manufacture can be included as a part of a computer system or sold separately.

[0042] Additionally, at least one program storage device readable by a machine embodying at least one program of instructions executable by the machine to perform the capabilities of the present invention can be provided.

[0043] The flow diagrams depicted herein are just examples. There may be many variations to these diagrams or the steps (or operations) described therein without departing from the spirit of the invention. For instance, the steps may be performed in a differing order, or steps may be added, deleted or modified. All of these variations are considered a part of the claimed invention.

[0044] Although preferred embodiments have been depicted and described in detail herein, it will be apparent to those skilled in the relevant art that various modifications, additions, substitutions and the like can be made without departing from the spirit of the invention and these are therefore considered to be within the scope of the invention as defined in the following claims.